Computational identification of regulatory regions and regulatory sites from high-throughput open chromatin assays

Galip Yardimci[1], Aslihan Karabacak[2], Antje Hirsekorn[2], Mahmoud Ibrahim[2], Michael Rauer[2], Christopher Frank[1], Gregory Crawford[1], Uwe Ohler[2]

[1]Center for Genomics and Computational Biology, Duke University, Durham NC, USA

[2]Berlin Institute for Medical Systems Biology, Max Delbruck Center Berlin, Germany

Approaches to map open chromatin have been shown to be highly useful to decode the regulatory genome as they provide information on gene regulation at multiple scales. For instance, mapping DNaseI hypersensitive regions (with a typical size of 200-300 nt) has been crucial to reduce the search space for functional regulatory interactions by 1-2 orders of magnitude, enabling us to develop predictive models of gene expression patterns or differentiation into different cell types (e.g. [1-2]). It has also been instrumental to our increased understanding of non-coding, pervasive transcription in the human genome and how it is encoded (e.g. [3]).

More controversial is the idea that at high, nucleotide-level resolution, these assays enable us to identify footprints of individual transcription factors. This would provide information on the precise interaction sites of many TFs simultaneously, and thus greatly reduce the need for factor—specific binding assays. However, DNase-seq data exhibits strong sequence bias, and reports have suggested that this renders it difficult if not impossible to identify genuine footprints.

We determined background cleavage preferences of DNaseI in naked, dechromatizined DNA from two ENCODE cell lines, and have developed a probabilistic mixture model for factor-specific footprint profiles that properly accounts for background while assessing the interaction evidence at each putative footprint [4]. We propose metrics and datasets that specifically allow to quantify the predictive contribution of the footprint alone, without conflating it with other sources of information such as accessibility in a wider region, and showed that results are informative for many but not all factors. As example application, we separate direct and indirect binding events in ChIP-seq data.

We will put these findings in the context of ongoing work using the more recent ATAC-seq protocol. This is an easier alternative to DNase-seq that requires less input material, but it has its own pitfalls and also shows extensive sequence bias. Based on our own fly and human cell line data sets, we will present our ATAC-seq processing strategy and discuss similarities and differences between ATAC-seq and DNase-seq.

References

[1] Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res 22:1711-22, 2012.

[2] Ibrahim MM, Lacadie SA, Ohler U. JAMM: a peak finder for joint analysis of NGS replicates. Bioinformatics 31:48-55, 2015.

[3] Duttke SH, Lacadie SA, Ibrahim MM, Glass CK, Corcoran DL, Benner C, Heinz S, Kadonaga JT, Ohler U. Human promoters are intrinsically directional. Mol Cell 57:674-84, 2015.

[4] Yardımcı GG, Frank CL, Crawford GE, Ohler U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Res. 42:11865-78, 2014.